Arkansas Law Review

Volume 72 | Number 2 Article 3

December 2019

Perversity as Rationality in Teacher Evaluation

Scott R. Bauries
University of Kentucky

Perversity as Rationality in Teacher Evaluation

Scott R. Bauries*

Introduction

Rational basis review is broken. Consider a vignette: Imagine a student, Lisa, who is about to graduate high school. Lisa has already completed all of the graduation course requirements early and is spending her time during her senior year taking interesting electives and dual-enrollment college courses. The state has a statute that requires school districts to deny a diploma to any student "who, during the final year of school attendance, fails to achieve a passing score on the state-approved, end-of-course exams in the courses of Language Arts, Mathematics, Science, and Social Studies in which that student is then-currently enrolled."

As part of the graduation requirements, schools must administer these "end-of-course" exams in the twelfth grade to every student enrolled in one of the aforementioned courses. It forbids early administration of the tests under any circumstances, due to concerns over cheating and test security. However, because Lisa took online courses in the summers, and completed her last graduation-required course in the eleventh grade, she took no end-of-course test then, and she will take no such test this year, as she is not enrolled in any graduation-required course. Thus, by operation of the mandate, Lisa will "fail to achieve a passing score" in all of the required subjects and will accordingly fail to graduate.

When Lisa and her parents notice this anomaly and inform the school district, the response is to quote the policy to them, and to express regret that Lisa will apparently not graduate.

---

* Associate Dean of Faculty Research and Willburt D. Ham Professor of Law, University of Kentucky. I would like to thank the University of Kentucky College of Law for financially supporting this research, and to thank the University of Arkansas School of Law and the *Arkansas Law Review* for inviting me to be a part of this important Symposium.

Obviously unsatisfied, Lisa's parents take Lisa's issue up the chain of command, all the way to the Superintendent's Office. The Superintendent, recognizing the absurdity of the situation, offers a solution. A random student from Lisa's graduating class will be selected, and Lisa's graduation requirement will be held satisfied if, and only if, *that* student passes all of the end-of-course exams.

Would that seem absurd or arbitrary? How about *irrational*?

Thankfully, no school district has such an arbitrary requirement, but what if one did? Would that pass muster in a challenge under the Fourteenth Amendment's Due Process Clause? Under current approaches, the answer would likely be *yes*. Focusing on a teacher evaluation plan in Florida, this contribution to the Symposium considers why this is, critiques that state of affairs, and offers the beginnings of a way forward, which, as it turns out, is somewhat a call for a way backward in Constitutional Law.

I. Teacher Evaluation Practices over Time

A. Historical Approaches to Teacher Evaluation

For most of the 350-year history of public education in the United States and the Colonies,[1] teachers were not evaluated for

1. Public education began in earnest with the Massachusetts Colony's "Old Deluder Satan Law" of 1647, which provided:

> It being one chief project of that old deluder, Satan, to keep men from the knowledge of the Scriptures, as in former times keeping them in an unknown tongue, so in these later times by perswading [sic] from the use of tongues, that so at least the true sense and meaning of the originall [sic] might be clowded [sic] by the false glosses of Saint-seeming deceivers; and that Learning may not be buried in the grave of our fore-fathers in Church and Commonwealth, the Lord assisting our endeavors: It is therefore ordered by this Court and Authoritie [sic] thereof; that every Township in this Jurisdiction, after the Lord hath increased them to the number of fifty Householders, shall then forthwith appoint one within their town to teach all such children as shall resort to him to write and read, whose wages shall be paid either by the Parents or Masters of such children, or by the Inhabitants in general, by way of supply, as the major part of those that order the prudentials of the Town shall appoint. Provided those which send their children be not oppressed by paying much more than they can have them taught for in other towns.

pedagogical performance or effectiveness.[2] Just before the turn of the Twentieth Century, policy makers and theorists became interested in evaluating teachers. Evaluative practices split into those favoring democratic participation and those favoring scientific approaches—the latter were the first to employ standardized tests and other results-based data, but only crudely.[3] During the period from the 1960s through the millennium, most teachers have been (and still are) evaluated, at least in part, through personal observations and rating rubrics, usually based on one or more class visits per year by an administrator or a fellow teacher.[4] This method of evaluating teachers was always subject to legitimate objections, as it is in the ordinary business context. Primarily, these objections centered upon bias, as the evaluations in question were usually the responsibility of one administrator.[5]

In the 1980s and 1990s, following the publication of the Reagan Administration's educational call-to-arms, *A Nation at Risk*,[6] policy makers became more interested in evaluating teachers as a means to improve schools, specifically through "merit pay," or pay-for-performance schemes.[7] In the late

> And it is further ordered, that where any town shall increase to the number of one hundred Families or Householders, they shall set up a Grammar-School, the Masters thereof being able to instruct youth so far as they may be fitted for the Universitie [sic]. And provided if any town neglect the performance hereof above one year then everie [sic] such town shall pay five pounds per annum to the next such School, till they shall perform this order.

Old Deluder Satan Law of 1647, https://www.mass.gov/files/documents/2016/08/ob/deludersatan.pdf?_ga=2.88969755.1272435236.1551892292-657769494.1551892292 [https://perma.cc/ST6N-D63V] (last visited Mar. 29, 2019); *see generally* Eric R. Ebeling, *Massachusetts Education Laws of 1642, 1647, and 1648, in* HIST. DICTIONARY AM. EDUC. 225, 225-26 (Richard J. Altenbaugh ed., 1999).

2. *See generally* ROBERT J. MARZANO ET AL., EFFECTIVE SUPERVISION: SUPPORTING THE ART AND SCIENCE OF TEACHING 12-29 (Deborah Siegel ed., 2011) (outlining the history of educational employee supervision and evaluation).

3. *Id*. at 14.

4. *Id*. at 28.

5. Arthur E. Wise et al., *Teacher Evaluation: A Study of Effective Practices*, 86 THE ELEMENTARY SCH. J. 60, 71 (1985).

6. NAT'L COMM'N ON EXCELLENCE IN EDUC., A NATION AT RISK: THE IMPERATIVE FOR EDUCATIONAL REFORM (1983).

7. *See* THERESA J. GURL ET. AL., POLICY, PROFESSIONALIZATION, PRIVATIZATION, AND PERFORMANCE ASSESSMENT 12 (2016); Wise et al., *supra* note 5, at 60.

1990s, and especially in the 21st Century, many states began using the standardized test scores of students, at least in part, to evaluate teachers on the theory that teachers should be accountable for students' results.[8] At its inception, there were many obvious problems with this idea. The existing tests were mostly designed for diagnostic—not evaluative—purposes, and few policy makers appropriately considered the many factors unrelated to teaching that might influence scores, such as poverty, family structure, race, family education levels, school safety, attendance, and the like.[9]

Measurement experts, psychologists, and statisticians worked for years to account for these problems so that teachers could be evaluated fairly based on their students' performance. Ultimately, they developed the highly controversial, but now widely used, technique of value-added modeling.[10]

B. Value-Added Modeling

Value-added modeling describes a group of highly complex statistical techniques that researchers and evaluators use to attempt to isolate the influence of an independent variable on the positive and negative changes in a dependent variable—in other words, to determine the "value" that the independent variable "adds" to the dependent variable.[11] When used for evaluating teachers, the independent variable is the performance effectiveness of the teacher in the classroom, and the dependent variable is the achievement of the evaluated teacher's students, as reflected in their standardized test scores.

Value-added models, such as the Florida model analyzed in the next section, use the prior performance of students on one-to-several years of standardized tests, among other factors, to

8. MARZANO ET AL., *supra* note 2, at 25.

9. *See* Edward Haertel, *The Valid Use of Student Performance Measures for Teacher Evaluation*, 8 EDUC. EVALUATION & POL'Y ANALYSIS 45, 46-50 (1986) (outlining the pitfalls of student test results for teacher evaluation).

10. *See* Douglas F. Warring, *Teacher Evaluations: Use or Misuse?*, 3 UNIVERSAL J. OF EDUC. RES. 703, 704 (2015) (situating the development of value-added modeling within the overall teacher evaluation debate).

11. *See id.* at 705 ("[V]alue-added model based on value added measures attempt to isolate the impact a teacher has on students' achievement from other factors of interest, such as student characteristics.").

compute an expected learning gain that each student should be able to accomplish in each subsequent testing year.[12] Then, for each student, the model computes the current-year test score that would be predicted based on one or more prior years of test scores, while attempting to control for student and school characteristics that are known to influence achievement, and compares the current–year test score actually obtained to that prediction to determine whether the actual score was higher or lower than what the model predicted it would be.[13]

As discussed above, the use of standardized test scores to evaluate teaching performance has always been controversial. Critics have objected to it for many reasons, including that standardized tests often measure only a narrow portion of what we hope students learn in school, and that they generally do so using the least expensive means—usually machine-scored multiple-choice questions—when other methods, such as essay or performance assessment, would be better aligned with the essential knowledge and skills we hope students will acquire in school.[14]

These problems, however, pale in comparison to the unfairness that results when student scores on standardized tests are directly imputed to schools and teachers as measures of educational quality, without controlling for other factors that may cause differences in scores.[15] Empirical research has established that, at most, between one and fourteen percent of the variation in student test scores can be attributed to the effectiveness of the teacher who taught the tested students in the

12. *See* Florida Value-Added Technical Report, Dkt. 86-2, Ex. 13C, at 2-3, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (N.D. Fla. 2014) [hereinafter Fla. Tech. Rep.].

13. OLIVIA LITTLE ET AL., A PRACTICAL GUIDE TO EVALUATING TEACHER EFFECTIVENESS 4 (Nat'l Ctr. For Tchr. Quality 2009); *see generally* LAURA GOE ET AL., APPROACHES TO EVALUATING TEACHER EFFECTIVENESS: A RESEARCH SYNTHESIS (Nat'l Comprehensive Ctr. For Tchr. Quality 2008) (discussing varying approaches to evaluating teacher effectiveness, including value-added).

14. Expert Report of Edward Henry Haertel, Dkt. 86-13, at ¶ 53, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (N.D. Fla. 2014) [hereinafter Haertel Rep.].

15. *See, e.g.*, W. James Popham, *Why Standardized Tests Don't Measure Educational Quality*, 56 EDUC. LEADERSHIP 8, 8-15 (1999) (outlining the inherent flaws in using standardized tests as measures of educational quality without controlling for other factors).

subject being evaluated.[16] This means that anywhere from 86 to 99 percent of the variation in student test scores is the result of factors other than the effectiveness of the teacher. Therefore, it would clearly be irrational to judge teaching performance, and then hold the teacher accountable for that performance, based on the test scores of a teacher's students without first controlling for all of those other causal factors, none of which lie within the evaluated teacher's control.

Statisticians and measurement experts are certainly not blind to this concern, nor are school leaders. Value-added modeling was initially conceived as a way of addressing this causation problem by statistically controlling for measured factors other than teaching performance, such as student prior performance, incoming language ability, socioeconomic status, race, and school characteristics, and thereby isolating the performance of a specific teacher as the cause of an identified learning gain.[17] Nevertheless, even with such controls, the use of value-added modeling remains quite controversial, mostly due to concerns over its validity and reliability.

"Validity" is a measurement term referring to "the usefulness of information that a test provides for decisions that need to be made."[18] In other words, concern over validity is a concern over the appropriateness of the inferences one seeks to draw, or the actions one seeks to take, based on the scores that a measurement yields. For example, even the best and most carefully calibrated weight scale will provide a poor—or invalid—measure of height. Weight and height are positively, but not perfectly, correlated. Thus, using a weight scale to measure height does provide *some* useful information, but to validly measure height, one needs a tool that is more directly reflective of height, such as a ruler or tape measure. Because many factors other than teaching influence student performance, some of which significantly, assessing validity in the context of value-added models is challenging. When many factors in

16. AMERICAN STATISTICAL ASSOCIATION, ASA STATEMENT ON USING VALUE-ADDED MODELS FOR EDUCATIONAL ASSESSMENT (2014), https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf [https://perma.cc/FF9Q-HXBF].

17. Haertel Rep., *supra* note 14, at ¶ 8.

18. Nancy Koh et al., *Understanding Validity Issues Surrounding Test-Based Accountability Measures in the US*, 22 QUALITY ASSURANCE IN EDUC. 42, 44-45 (2014).

addition to the construct the evaluator is interested in affect a variable, the factors other than the construct need to be "controlled."[19]

The main purpose of value-added modeling is therefore to isolate the teaching performance of a particular teacher in driving student achievement by controlling for measurable factors other than teaching that contribute to student achievement.[20] But some factors are very difficult to control for.[21] Among these are the validity and/or reliability of the underlying standardized tests used to judge student achievement, nonrandom assignment of students to teachers,[22] and summer learning loss, if a model uses a prior year's assessment as a pre-test.[23] Thus, even the most carefully designed value-added models suffer from validity concerns.[24]

Aside from concerns over validity, which can be minimized (though never eliminated) through careful model design and implementation, another problem with value-added models is their very low "reliability."[25] Reliability is a measurement term used to describe the consistency of a test in measuring the same construct from one administration to the next.[26] For example, though a weight scale is obviously a *valid* measure of weight, for the scale to be a *reliable* measure of weight, it must read "10 pounds" when a 10-pound weight is placed on it, and it must do so every time the same weight is placed on it. If it reads "10 pounds" the first time, and then "4 pounds" the second time for the same object, then the scale is not a reliable measure of weight.

---

19. Matthew Johnson et al., *Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables*, 2 (Mathematica Policy Research, Working Paper No. 25, 2013).

20. *Id*. at 2.

21. Preston C. Green et al., *The Legal and Policy Implications of Value-Added Teacher Assessment Policies*, 2012 BYU EDUC. & L. J. 1, 6 (2012).

22. LITTLE ET AL., *supra* note 13, at 6.

23. Haertel Rep., *supra* note 14, at ¶ 25.

24. LITTLE ET AL., *supra* note 13, at 5.

25. *See* Green et al., *supra* note 21, at 6-7; Haertel Rep., *supra* note 14, at ¶ 44.

26. Green et al., *supra* note 21, at 6.

In measurement scholarship, reliability is measured using a coefficient, the value of which can range from zero to one.[27] A value of one indicates perfect reliability—a scale that reads "10 pounds" every time the 10-pound bowling ball is placed on it. A score of zero indicates no reliability—a scale that might read literally *any* value each time the same 10-pound bowling ball is placed on it.

Scholarship has established that the reliability of value-added model scores from year to year ranges between .2 to .3—or what would be considered very low reliability—not much better than chance.[28] By way of comparison, well-known standardized tests such as the SAT and the ACT typically have reliability coefficients on the order of .8 to .9.[29] It would not be rational for a decision maker seeking to come to a defensible decision—especially an important one—to rely on an instrument with very low reliability. In fact, as a recent study put it, "[c]oefficients at or above 0.80 are often considered sufficiently reliable to make decisions about individuals based on their observed scores, although a higher value, perhaps 0.90, is preferred if the decisions have significant consequences."[30] Because value-added modeling is used to evaluate teachers for the purposes of promotion, tenure, and potentially even dismissal, reliability is a major concern.

To better understand how weak these value-added model reliability coefficients are, the authors of one study divided the teachers evaluated into quintiles and tracked the stability of their

27. *See* Noreen M. Webb et al., *Reliability Coefficients and Generalizability Theory in* HANDBOOK OF STATISTICS, VOL. 26, 81-120 (C.R. Rao & S. Sinharay eds., 2007) (explaining reliability coefficients and their purposes).

28. *See* Green et al., *supra* note 21, at 7 (the year-to-year correlation is .2 to .3); Daniel F. McCaffrey et al., *The Intertemporal Variability of Teacher Effect Estimates*, 4 EDUC. FIN & POL'Y 572, 588 (2009) (year-to-year correlations range from 0.2 to 0.5 for elementary school and 0.3-0.7 for middle school).

29. *See,* e.g., ACT, INC., *The ACT Technical Manual* tbl.10.2.1 (2017), http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf [https://perma.cc/Z8J6-2N99]; *see also* THE COLLEGE BOARD, *Test Characteristics of the SAT: Reliability, Difficulty Levels, Completion Rates* 1 (2013), http://media.collegeboard.com/digitalServices/pdf/research/Test-Characteristics-of-SAT-2013.pdf [https://perma.cc/32V7-BXCG].

30. Webb et al., *supra* note 27, at 81.

ratings placement from quintile to quintile over two years.[31] They found that teachers who scored in the top quintile one year were just as likely to find themselves in one of the bottom two quintiles the next year as they were to find themselves in the top quintile again.[32] Such large changes in ratings from year to year indicate that the value-added model studied (a precursor to Florida's current model, discussed in the next section) was very imprecise, or unreliable.[33] Such instability alone should cause concern in legislatures considering using value-added modeling for high-stakes decisions, especially where (as is true three states) the model accounts for half of the teacher's rated effectiveness.[34]

Despite these flaws, it is accepted by a portion of the scholarly community that, when meticulously constructed and used as designed, a value-added model can provide enough useful information over time to justify using model-derived ratings, but only as one element among others of the overall evaluation of a teacher. Much skepticism remains over uses, such as the one described below, that assign 50 percent of a teacher's effectiveness rating to a value-added model, and then attach significant consequences to that rating.[35] Because value-added models are often used to evaluate individual teachers for job-related benefits and consequences, critical to defending such uses is designing a model that clearly and unambiguously isolates one teacher's influence on student learning.[36] As discussed below, Florida's most recent value-added assessment program as applied in four individual school districts not only

31. McCaffrey et al., *supra* note 28, at 574 (citing Cory Koedel & Julian R. Betts, *Re-Examining the Role of Teacher Quality in the Educational Production Function* (Univ. of Missouri-Columbia, Working Paper No. 708, 2007)).

32. *Id*. at 574-75.

33. Green et al., *supra* note 21, at 6-7.

34. *See id*. at 3-5 (describing three state programs that base 50 percent or more of the teacher rating on value-added models).

35. *Id*. at 21-22.

36. *See* ERIN D. LOMAX & JEFFREY J. KUENZI., CONG. RESEARCH SERV., R4105, VALUE-ADDED MODELING FOR TEACHER EFFECTIVENESS 4 (2012) ("VAM recognizes that there are multiple factors that contribute to learning and is therefore designed with the intention of isolating the teacher's effect on student learning. The 'teacher effect' is an estimate of the teacher's unique contribution to student achievement as measured by student performance on assessments.").

fell short of, but also actively worked against, this critical consideration that underlies all value-added modeling.

II. Florida's Experiment with Value-Added Modeling

This section reviews the recent effort of teacher groups to challenge the teacher evaluation system in Florida, as applied in four public school districts. This unsuccessful challenge illustrates the perverse policy incentives and results that the modern approach to rational basis review yields.

A. Value-Added Modeling in Florida

Like several other states, mostly in response to an Obama-era competitive federal funding program called Race to the Top,[37] Florida opted to evaluate its public school teachers using value-added modeling. Under the legislation requiring this form of evaluation, at least one-third of a teacher's "effectiveness" score, which has implications for retention, remediation, and salary increases, must be based on a value-added model score of that teacher's effectiveness.[38]

Like all teacher evaluation systems that employ value-added modeling, Florida's system is designed to isolate one teacher's influence on the testing performance of that teacher's students.[39] It does so by collecting the students' scores on the Florida Comprehensive Assessment Test (the "FCAT") in reading in grades 3 through 10 and mathematics in grades 3 through 8, and computing a score for each student that reflects the difference between what that student's prior performance would have predicted, and what the student actually achieved.[40] Once all of these scores are computed, they are combined with each other, and a series of statistical controls are then applied to account for non-teaching factors that could have influenced learning gains or losses.[41] The resulting score is then further

37. *See, e.g.*, Green et al., *supra* note 21, at 1 n.2.

38. FLA. STAT. § 1012.34(3)(a)(1); *see also* Haertel Rep., *supra* note 14, at ¶ 11.

39. Fla. Tech. Rep., *supra* note 12, at 1.

40. *See* Cook v. Stewart, 28 F. Supp. 3d 1207, 1210 (N.D. Fla. 2014) (describing the program's details as part of a challenge to its constitutionality).

41. Fla. Tech. Rep., *supra* note 12, at 3-4. According to the State's technical report, these factors include each student's prior test scores; the number of courses in the tested

controlled for the overall achievement levels of the school's student body, in a laudable effort to capture societal and community effects.[42]

Once the model controls for these factors, a large portion of the variance in test scores, and therefore a large portion of the gains or losses in achievement among the teacher's students, will have been accounted for by the controlled non-teaching factors. Importantly, *all* the residual portion of the student learning gains *not* accounted for by these controlled factors is then assumed to be caused by the teaching performance of the teacher who taught the tested students in the tested subject.[43] In other words, even when used as designed, the Florida Value-Added Model does not arrive at a direct conclusion, but an indirect one based on the existence of a residual student gain not accounted for by the controlled non-teaching factors.[44] Put differently, it assumes that student achievement not caused by the non-teaching factors specified in the model was caused entirely by the individual teacher's effective or ineffective performance.

Such an assumption depends heavily on the further assumption that all relevant non-teacher-specific factors have been adequately accounted for in the model, but some outside factors are impossible to control in any statistical model of this type.[45] For example, as outlined above, scholars have identified "peer effects"—the increases and decreases in learning growth that a student experiences by being placed in classes with strong or weak students—as a factor that value-added modeling has trouble controlling for.[46] Additionally, in Florida, because the value-added model compares the scores of students at the end of

---

subject that each student takes; each student's disabilities (if any); each student's ability to speak and read English; whether each student is gifted; each student's attendance record; the mobility of each student from school to school during the school year; the tendency of students to be promoted to the next grade after one year; the size of the class each student is in for the tested subject; and the existing differences, or variance, in test scores among the students in the tested class. *Id*.

42. *Id*. at 4-6.

43. *Id*. at 6-7.

44. *Id*.

45. Haertel Rep., *supra* note 14, at ¶ 23.

46. NAT'L RESEARCH COUNCIL & NAT'L ACAD. OF EDUC., GETTING VALUE OUT OF VALUE ADDED: REPORT OF A WORKSHOP 46 (Henry Braun, Naomi Chudowsky & Judith Koenig eds., 2010).

one academic year to the scores of those same students at the end of another academic year, it cannot control for learning losses that occur over the summer, and this failure to control has a disparate impact based on both race and socioeconomic status.[47]

In the absence of an adequate control for summer learning loss, for example, a teacher of reading or math will be held responsible for such loss, even though she lacks any ability to prevent it because she will not even meet the tested students in question until after the learning loss happens. In short, there is a significant concern that, even when used as designed, value-added models such as Florida's may be measuring the influence of factors confounded with the factor they attempt to isolate—teacher effectiveness. In measurement terms, value-added models such as the one used by Florida therefore may not be valid measures of teaching quality.[48]

Two other major problems affect value-added modeling in general. One is the problem of what scholars term "spillover effects"—the unmeasurable, but real, effects that a team of teachers teaching the same students can have on each other's students.[49] The other is the impossibility of deriving any sort of value-added rating for teachers who either do not teach the tested material, or do not teach any material in a tested grade.[50] These problems existed in Florida's model, as they do to a certain extent in all value-added evaluation models. Efforts to preserve the high-stakes use of value-added modeling for all public school teachers at all costs, however, led Florida to approve uses of the model directly in conflict with its purposes, methods, and specifications, and, in one case, directly in conflict with each other. These approved uses led to a judicial challenge, which the next section explicates.

---

47. Haertel Rep., *supra* note 14, at ¶ 25.

48. Under Eleventh Circuit precedent, this concern alone arguably should have justified invalidation of the Model. *See* Debra P. v. Turlington, 644 F.2d 397, 404-06 (5th Cir. 1981) (remanding for a showing that what was tested on a high school exit examination was actually taught in Florida's high schools).

49. *See* Green et al., *supra* note 21, at 10 (discussing "spillover effects").

50. *Id*. at 14-15 (identifying this problem).

B. Evaluating Teacher Evaluation in Florida

In *Cook v. Stewart*,[51] later styled on appeal *Cook v. Bennett*,[52] a group of plaintiffs challenged the public school teacher evaluation system outlined above, as applied in three counties of the State. These three counties, with the State's approval, chose to account for the lack of usable scores for teachers in some grade levels and subjects by attributing to those teachers the test performance of students whom the evaluated teachers either did not teach at all, or did not teach the tested subjects.[53]

The District Court in *Stewart* entered judgment for the State and the District defendants, based on two separate orders.[54] Both of these orders held that the decision makers for each defendant "could rationally believe" that the use of value-added ratings computed from the test scores of one teacher's students to assign a performance rating to another teacher who did not teach those students, and/or did not teach the tested subject, furthered a state interest in improving student achievement.[55] The remainder of this Part evaluates these conclusions.

1. Legislating or Mandating Internal Contradictions

The challenges to the programs in Florida sounded in both substantive due process and equal protection, and accordingly, the court applied rational basis review to evaluate their constitutionality.[56] At a minimum, it would seem, even under current approaches to rationality review of legislation, a law should be invalidated if the means adopted to serve a legitimate end are more likely to frustrate than serve the end, or if the means are developed by experts and are put to a use that is

51. Cook v. Stewart, 28 F. Supp. 3d 1207 (N.D. Fla. 2014).

52. Cook v. Bennett, 792 F.3d 1294 (11th Cir. 2015).

53. *Stewart*, 28 F. Supp. 3d at 1212.

54. *See generally* Order Granting in Part and Denying in Part State Defendant's Motion to Dismiss at 17, *Stewart*, 28 F. Supp. 3d 1207 (No. 1:13-cv-72-MW-GRJ); Order on Cross-Motions for Summary Judgment at 17, *Stewart*, 28 F. Supp. 3d 1207 (No. 1:13-cv-72-MW-GRJ).

55. *Stewart*, 28 F. Supp. 3d at 1213.

56. *Id*. at 1212-14.

directly contrary to the assumptions on which the experts designed it.[57] But this was not the case in Florida.

As introduced above, value-added modeling attempts to address the many objections to the use of standardized test scores to judge educational quality by isolating the effect of one teacher on the standardized test scores of that teacher's students, in the subject or subjects that teacher teaches, while controlling for the influence of other factors on such scores. Also as outlined above, the assumption of the value-added model used in Florida is that, once all of the non-teaching factors are controlled for, *all* of the remaining non-random variation in student achievement on the FCAT is attributable to the efficacy of the student's teacher with respect to the tested curriculum. So, the Florida model's design requires (1) accounting for all measurable factors that might explain student performance, other than the performance of the student's teacher in the tested curriculum; (2) assuming that *all* variation in student scores not explained by those non-teaching factors was caused by the student's teacher in the tested curriculum; and then (3) computing a rating for that teacher based on that residual portion of student score gains, adjusted for overall student achievement in the teacher's school.[58] If one takes this model seriously, then *no* additional causes of student achievement are possible.

Florida's model takes its student test scores from the annually administered FCAT, which tests reading in grades 3 through 10, and math in grades 3 through 8.[59] No other scores were used in the three defendant districts, so teacher evaluators were presented with the two problems introduced above.[60] One was how to assign ratings to teachers who teach in tested grades,

57. This was, in fact, arguably the controlling law on the books at the time in the 11th Circuit, as reflected in the substantive due process case, Debra P. v. Turlington, 644 F.2d 397, 404 (5th Cir. 1981), in which the court explained that "the state is obligated to avoid action which is arbitrary and capricious, does not achieve or even frustrates a legitimate state interest, or is fundamentally unfair." The Eleventh Circuit, in Bonner v. City of Prichard, 661 F.2d 1206, 1209 (11th Cir. 1981), adopted as controlling precedent the decisions of the 5th Circuit "as that court existed on September 30, 1981, handed down by that court prior to close of business on that date." *Debra P.* was decided on May 4, 1981, and rehearing denied on September 4, 1981, *Debra P.*, 644 F.2d at 397, so it is controlling 11th Circuit precedent, but the case is not mentioned in *Stewart*.

58. Fla. Tech. Rep., *supra* note 12, at 2-7.

59. *Id*. at 1.

60. Haertel Rep., *supra* note 14, at ¶ 15-16.

but do not teach reading or math.[61] The other was how to assign ratings to teachers who do not teach any subject in one of the tested grade levels.[62]

To evaluate the uses to which Florida and the three districts put the Florida model, and to illustrate how the school districts, with the State's approval, chose to address these problems, the District Court divided the plaintiff teachers into two groups, based on the circumstances that caused them to object to the use of these computed scores to judge their teaching.[63] What the District Court termed "Type B" teachers were those who taught students in grades in which the FCAT was administered, but who did not teach any FCAT-tested curriculum to those students—a seventh grade music teacher, for example.[64] What the District Court termed "Type C" teachers were those who taught in grades in which no students took the FCAT—kindergarten through second grade, as well as eleventh and twelfth grade—or in third grade, the year students take only the baseline (first administration) test, thereby making the computation of any student growth score impossible.[65]

The districts computed value-added ratings for the Type B teachers based on the reading FCAT scores of the students whom the Type B teachers taught non-tested curricula, such as music or science.[66] For the Type C teachers, even this was not possible, so the districts assigned each of these teachers a value-added rating made up entirely of the portion of the variance in test scores attributable to non-teaching factors at the teachers' schools.[67] Indeed, one of the school district defendants, Alachua County Schools, even evaluated the teachers of *one* elementary school that contained only grades kindergarten through second

61. *Id*. at ¶ 16.
62. *Id*. at ¶ 15.
63. Cook v. Stewart, 28 F. Supp. 3d 1207, 1210 (N.D. Fla. 2014).
64. *Id*.
65. *Id*.
66. Haertel Rep., *supra* note 14, at ¶ 16.
67. *Id*. at ¶ 15. As Dr. Haertel, the Appellants' expert, explains, in two of the Districts, the score was actually a combination of the school portion and the average of the teacher portion for all of the teachers in the school, but since the teachers' value-added scores would have naturally roughly balanced each other out, the scores in these Districts were actually nearly entirely a reflection of the school portion. *Id*. at ¶ 15, n.2.

using the schoolwide test scores of the fourth and fifth grade students of a *completely different* elementary school.[68]

It should be readily apparent that these uses were completely antithetical to the methodological purpose of value-added modeling, which at a minimum, seeks to isolate one teacher's influence on student performance. As to the Type B teachers, assigning a value-added rating to a teacher who did not teach the curriculum tested, while also attributing *that same* residual variation in student scores after controlling for non-teaching factors to the teacher who actually *did* teach the tested curriculum, contradicts the model's specifications directly. The model, recall, assumes that 100 percent of any residual variation in student scores left once all control factors are accounted for is caused by the teacher who taught the tested students in the tested subject.[69] But the approach the districts took with the Type B teachers also attributed *100 percent of that same residual variance* to every Type B teacher in the school who taught the same students.

Thus, the districts' use of the model with Type B teachers directly contradicted both the purpose of the model—to isolate one teacher's influence on student achievement—and the model's specifications. But worse than this, it had the effect of holding one teacher responsible for the classroom performance of another teacher not subject to the Type B teacher's supervision or control. Even assuming the existence of a "spillover effect"[70] that causes achievement effects across a grade-level teaching team, any such effects were not subject to the direct control of the Type B teacher and were therefore an arbitrary means of rating that Type B teacher.

As to the Type C teachers, the use of the portion of student score variance explained by non-teaching school factors identified a covariate designed to control for school characteristics, and to thereby make the individual teacher ratings more accurate and valid by adjusting for between-school differences, and instead used it as the sole determinant of

68. Plaintiff's Motion for Summary Judgment and Incorporated Memorandum of Law at 2, Cook v. Stewart, 28 F. Supp. 3d 1207 (N.D. Fla. 2014) (No. 1:13-cv-00072-MW-GRJ).

69. Fla. Tech. Rep., *supra* note 12, at 6.

70. *See* Green et al., *supra* note 21, at 6 (discussing spillover effects).

whether a Type C teacher was performing well. In other words, even though the central purpose of using a value-added model to evaluate teaching employees is to isolate the influence of one teacher's performance on her student's achievement, the ratings of the Type C teachers isolated precisely nothing at the teacher level.

Type C teachers were instead rated based on the overall performance of students in the school who took the FCAT. No attempt was made to isolate any influence that the Type C teacher—or indeed, any other teacher—had on that performance. Rather than separating effective from ineffective teachers, then, the model as applied to Type C teachers rated every single teacher in the same school who did not teach FCAT-tested students or subjects as equally effective or ineffective. Such a use, like the use to which the model was put with the Type B teachers, was directly in conflict with both the model's purpose and design.

The State's ostensible goal in adopting value-added modeling as the basis for teacher evaluation statewide was "increasing student success" (from the State's summary judgment brief),[71] or "increasing student learning growth" (from the District Court's opinion).[72] However, the State's goal would have been meaningless in the context of value-added assessment unless the assumption underlying it was that, when teachers receive lower value-added scores, they will respond to those scores by taking action to improve their practices, thereby improving student achievement and increasing their own value-added ratings, in the hopes of both improving their practice and avoiding negative consequences, such as dismissal. Basing the value-added score for the Type B and Type C teachers on the performance of students they either did not teach at all, or did not teach the tested curriculum, instead based the score entirely on matters that were outside the direct control or influence of the teachers, leaving these teachers no way to respond to a bad score to improve the achievement of the tested students.

The logic of the value-added system itself would contradict this use. Recall that, under the State's value-added model, *all* of

71. Defendant's Motion for Summary Judgment at 19, *Stewart*, 28 F. Supp. 3d 1207 (No. 1:13-cv-00072-MW-GRJ).

72. *Stewart*, 28 F. Supp. 3d at 1212.

the residual score variation left over after the non-teaching factors are accounted for was attributed to the student's teacher in the tested subject, logically meaning that, if we were to believe the model, *none* of this residual variation was attributable to any other cause, including the performance of any other teacher. Were that not the case, it would have been irrational to rate the teacher of the tested students in their tested subject based on that residual—that's the entire purpose of value-added modeling, to control for factors other than the rated teacher's performance.[73] Similarly, recall that the school's overall score was not connected to any particular teacher, but was the State's way of calibrating overall student achievement levels in the school in the tested subjects to account for the differences between schools as a control variable. So, neither of these outcomes were subject to the influence or the control of any Type B or Type C teacher. In short, there was literally nothing any Type B or Type C teacher could have done purposely to improve their own teaching in response to their value-added ratings, because neither the Type B ratings nor the Type C ratings contained any useful information about these teachers' own teaching performance.

To illustrate, under the uses of the value-added model adopted by the districts and approved by the State, if an ineffective Type B teacher were lucky enough to share students with an *exemplary* reading teacher, for example, that Type B teacher would be judged to be an excellent performer based on that exemplary reading teacher's good performance, despite the Type B teacher's own possibly *ineffective* teaching of his or her own subject. Conversely, if a highly *effective* Type B teacher were unlucky enough to share students with a particularly *poor* reading teacher, the Type B teacher would be judged to be a substandard teacher, despite that Type B teacher's own possibly *excellent* teaching performance. The only thing that the Type B teacher would be able do in such a case would be to work the back channels of her school administration to make sure that she does not share any students with the poor reading teacher the next year. This outcome manifestly would not serve the purpose of "increasing student success" or "increasing student learning

73. *See* Johnson et al., *supra* note 19, at 2.

growth." In fact, because it would incentivize not better teaching but administrative gamesmanship, it would patently work *against* that purpose.

Similar to the Type B teachers, the Type C teachers, who were rated based on the overall aggregate performance of the students in their schools, could not control or change the characteristics of the schools into which they were assigned, and they could not do anything to influence, for example, the quality of the principal's leadership, the school's faculty-student ratio, or the average years of experience of the teachers with whom they taught—all non-teaching factors that might plausibly be factors influencing the overall school score. So, if, for example, a Type C teacher who was an *exemplary* classroom teacher were recruited to a struggling school to teach disadvantaged students, and she did a terrific job with *her own* students, but she did not teach any FCAT-tested grade levels, she would nevertheless be rated as a poor teacher if the FCAT scores of the students she did not teach in the *other grades* were to fall short of their predicted growth. The only thing that such a teacher would be able to do in such circumstances would be to work the back channels of administration to secure an assignment to a more advantaged school. Once again, since it would incentivize administrative gamesmanship rather than better teaching, this outcome would be manifestly at odds with the ostensible state goal of "increasing student success" or "increasing student learning growth."[74]

The District Court's opinion elided these obvious problems and judged to be "rational" a severely attenuated—one might say fanciful—theory of causation that was squarely at odds with both the purpose of value-added modeling and the evidence in the record.[75] This theory of causation held that the defendants "could rationally believe" that, by contributing positively or negatively to the overall learning environment of the school, each teacher in a school would have effects on the performance of their own students and that of other students in the school in subjects and grades the teacher did not teach.[76] It was therefore

74. Defendant's Motion for Summary Judgment, *supra* note 71, at 19; *Stewart*, 28 F. Supp. 3d at 1212.

75. *Stewart*, 28 F. Supp. 3d at 1213.

76. *Id*.

rational for the defendants to rate these Type B and Type C teachers' effectiveness based on student performance in subjects and/or grades these teachers did not teach.[77]

Thus, the court's theory of causation, which the State defendants offered, but which the court devised on behalf of the districts (which had not even moved for summary judgment themselves), converted the theorized "spillover effect," a confounding factor for which value-added modeling is supposed to control, into the independent variable in the analysis for the Type B teachers.[78] And it converted overall school-level student achievement, another confounding factor for which the model was supposed to control, into the independent variable for the Type C teachers.[79]

Under the District Court's reasoning, if the teacher evaluation systems in the Districts were instead based on increases and decreases in sales of healthy food in the school lunchroom (either to the teacher's own students or to the student body as a whole), then it would be "rational" to hire, fire, tenure, deny tenure to, or otherwise discipline the teachers based on those sales because it is conceivable that one could rationally believe that all teachers in a school should be promoting healthy lifestyles, and that the healthier a student's eating choices are, the more likely that student will be ready, willing, and able to learn—thereby improving student achievement. Obviously, using such a method for rating teachers would be ridiculous, but the Type B and Type C teachers had no more control over the teaching of their colleagues in other grades and/or subjects than they did over the sales abilities of the cafeteria staff in their schools.

Considering the State's purported justification for its value-added model of improving student achievement by holding teachers accountable for the test results they produce, it is impossible to square the methods described above with that goal. In fact, by holding teachers accountable for performance they can influence only incidentally, if at all, the model as applied to the Type B and Type C teachers worked directly against this goal. Moreover, the model adopted in the

77. *Id*. at 1212-14.
78. *Id*.
79. *Id*. at 1211-15.

defendants' districts, which was adapted from a state-level model carefully constructed by experts, directly violated the assumptions these experts used to construct their model. At a minimum, even under the current, very deferential approach to rational basis review, adopting means that directly frustrate one's stated goals and underlying assumptions should have been a bridge too far.

2. Legislating without Seeking Objective Expert Information

More controversially, earlier approaches to rationality review, such as those employed in the *Lochner* era, would not only seek to know whether the means adopted to serve a legitimate legislative end would instead frustrate that end, but also whether the legislature had established, as a factual matter, the need for the means it had chosen, and the effectiveness of the chosen means at meeting the need. Based on the majority opinion in *Lochner*, the failure of the New York legislature to do so was what doomed its maximum hours law.[80] But here again, against the challenge that the Florida legislature's chosen use of value-added modeling to judge one teacher's effectiveness based on the scores of another teacher's students lacked support, the court upheld the law.[81]

Measurement scholarship has established that between one and fourteen percent of a student's standardized test score gains can be attributed to the effectiveness of the student's teacher in the tested subject based on value-added modeling, and that only where careful controls are placed on the model.[82] However, no scholarship whatsoever has established that any portion of a student's test score performance can be isolated and explained by the teaching performance of teachers who do not teach that student, or who do not teach the tested curriculum.

This lack of scholarly support is not surprising. The uses to which the districts put Florida's value-added model were directly in conflict with the purpose of value-added modeling.[83]

---

80. Lochner v. New York, 198 U.S. 45, 62-64 (1905).
81. Cook v. Stewart, 28 F. Supp. 3d 1207, 1212-14 (N.D. Fla. 2014).
82. AMERICAN STATISTICAL ASSOCIATION, *supra* note 16.
83. Haertel Rep., *supra* note 14, at ¶ 57.

One need not be a measurement expert to understand why there is a complete lack of any scholarship even hinting at examining the hypothesis that one employee's performance can be assessed based on the performance of a completely different employee over whom the evaluated employee has no control or authority. Scholars do not study these methods for the same reason they do not evaluate whether to award a diploma to one student based on another student's standardized test scores, as in the vignette at the beginning of this article—it is facially preposterous and patently irrational to even consider doing that. Likewise, it is preposterous to believe that one teacher can or should be held accountable for the growth or lack thereof in test scores of students they do not teach, or on tests given to assess a curriculum they do not teach, and for which they do not claim any expertise.

No rational school district would voluntarily adopt such a system, no rational parents would choose to have their children's teachers evaluated in this manner, and no rational teacher would choose to be evaluated in this way. The District Court even said as much in the conclusion to its Order granting summary to the State and districts:

> The unfairness of the evaluation system as implemented is not lost on this Court. We have a teacher evaluation system in Florida that is supposed to measure the individual effectiveness of each teacher. But as the Plaintiffs have shown, the standards for evaluation differ significantly. FCAT teachers are being evaluated using an FCAT VAM that provides an individual measurement of a teacher's contribution to student improvement in the subjects they teach. The FCAT VAM has been applied to Type B teachers as well, but perversely it can only measure student improvement in subjects not taught by the Type B teacher. For Type C teachers the FCAT VAM has been applied as a school-wide composite score that is the same for every teacher in the school. It does not contain any measure of student learning growth of the Type C teacher's own students. To make matters worse, the legislature has mandated that teacher ratings be used to make important employment decisions such as pay, promotion, assignment, and retention. Ratings affect a teacher's professional

> reputation as well because they are made public—they have even been printed in the newspaper. Needless to say, this Court would be hard-pressed to find anyone who would find this evaluation system fair to non-FCAT teachers, let alone be willing to submit to a similar evaluation system.[84]

But in the next breath, the court stated, "[f]or reasons that have been explained, the State Defendants could rationally conclude that the evaluation policies further the state's legitimate interest in increasing student learning growth."[85]

To recap, despite acknowledging that it would be difficult to find a person who would themselves be willing to be evaluated in this way, the court felt compelled to find that two evaluation programs, one of which rated the performance of teachers based on the test scores students they did not teach the tested material, and the other of which rated their performance based on the test scores of students they did not teach at all, were rationally related to the legitimate government interest of improving student achievement. Both of these rating plans stood in direct conflict with both the purpose of the program and the specifications of the model derived through hours of professional work. Despite the fact that neither of these groups of rated teachers received any information from the ratings that they could individually use to improve their own teaching practice, the courts held that the government "could rationally believe" that so rating these teachers, and then attaching potentially extreme consequences to the ratings, would somehow improve student achievement. And under the law today, these decisions were likely *correct*. So, why did the District Court—and the Court of Appeals, which affirmed the rulings in all respects[86]— feel compelled to conclude that the program was rationally related to a legitimate government interest?

III. Rational Basis Review and the *Lochner* Recoil

The answer is the modern approach to rational basis review. Review of legislation for whether it bears a rational

84. *Stewart*, 28 F. Supp. 3d at 1215-16.
85. *Id*. at 1216.
86. Cook v. Bennett, 792 F.3d 1294, 1294 (2015).

relationship to a legitimate legislative interest in the public health, safety, welfare, or morals has been a feature of constitutional law for a very long time. The traditional approach to review of legislation under the Due Process Clause of the Fourteenth Amendment—what is generally referred to as "substantive due process"—is exemplified by the cases decided in what is now known as the "*Lochner* Era."

Lochner v. New York,[87] a case so infamous as to have a place in the "anticanon"[88] alongside such judicial embarrassments as *Korematsu v. United States*[89] and *Dred Scott v. Sandford*,[90] invalidated a New York law limiting the number of hours a baker could be required to work to no more than sixty per week.[91] Today, the case is nearly universally reviled, chiefly due to its recognition of a right to contract to sell one's labor as a liberty interest under the Due Process Clause of the Fourteenth Amendment. This recognition—which stemmed from cases preceding *Lochner*, but was forcefully applied in *Lochner*—allowed, or at that time required, the Court to examine the law as a valid exercise of the police power of New York to regulate the health, safety, morals, and welfare of its citizens.[92]

In conducting its review, the Court considered the State's proffered justification that regulating the hours of weekly work for a baker was an exercise of the power to regulate the public health, the argument being that extended work in baking exposes workers to a higher risk of respiratory ailments.[93] Stating its role in reviewing such legislation where the liberty interest in question is the liberty to enter into labor contracts, the Court set forth that era's version of rational basis review:

> It must, of course, be conceded that there is a limit to the valid exercise of the police power by the State. There is no dispute concerning this general proposition. Otherwise the Fourteenth Amendment would have no efficacy and the legislatures of the States would have unbounded power, and

87. Lochner v. New York, 198 U.S. 45, 57-64 (1905).

88. For a detailed and careful treatment of the "anticanon," which includes *Lochner*, see Jamal Greene, *The Anticanon*, 125 HARV. L. REV. 379 (2011).

89. 323 U.S. 214 (1944).

90. 60 U.S. 393 (1857).

91. *Lochner*, 198 U.S. at 64.

92. *Id*.

93. *Id*. at 50-51.

> it would be enough to say that any piece of legislation was enacted to conserve the morals, the health or the safety of the people; such legislation would be valid, no matter how absolutely without foundation the claim might be. The claim of the police power would be a mere pretext — become another and delusive name for the supreme sovereignty of the State to be exercised free from constitutional restraint. This is not contended for. *In every case that comes before this court, therefore, where legislation of this character is concerned and where the protection of the Federal Constitution is sought, the question necessarily arises: Is this a fair, reasonable and appropriate exercise of the police power of the State, or is it an unreasonable, unnecessary and arbitrary interference with the right of the individual to his personal liberty or to enter into those contracts in relation to labor which may seem to him appropriate or necessary for the support of himself and his family?* Of course the liberty of contract relating to labor includes both parties to it. The one has as much right to purchase as the other to sell labor.[94]

The Court did not mention anything about "fundamental" rights,[95] even though other courts of the era sometimes used that adjective.[96] Following this statement, the Court analyzed the State's health-based justification and found a factual foundation for it lacking:

> The act is not, within any fair meaning of the term, a health law, but is an illegal interference with the rights of individuals, both employers and employes [*sic*], to make contracts regarding labor upon such terms as they may think best, or which they may agree upon with the other parties to such contracts. Statutes of the nature of that under review, limiting the hours in which grown and intelligent men may labor to earn their living, are mere meddlesome interferences with the rights of the individual, and they are not saved from condemnation by the claim that they are

94. *Id*. at 56 (emphasis added).

95. Lochner v. New York, 198 U.S. 45 (1905).

96. *E.g.*, Meyer v. Nebraska, 262 U.S. 390, 401 (1923) ("That the State may do much, go very far, indeed, in order to improve the quality of its citizens, physically, mentally and morally, is clear; but the individual has certain fundamental rights which must be respected.").

> passed in the exercise of the police power and upon the subject of the health of the individual whose rights are interfered with, *unless there be some fair ground, reasonable in and of itself, to say that there is material danger to the public health or to the health of the employes* [*sic*], *if the hours of labor are not curtailed*.[97]

The emphasized portion above sets up the means-ends scrutiny that is now familiar to any student of constitutional law, but it sets up a particularly searching and skeptical version of it, requiring the establishment of an objectively reasonable concern that requires a legislative remedy of the type the legislature has chosen—a concern that actually motivated the legislature's choice.[98]

Rhetorically, at least, this test is not as far removed from current approaches as the case's status in the "anticanon" would indicate. Nevertheless, it was not long before the Court chose to abandon review of statutes for substantive reasonable necessity, documented by facts presented and proved by the state defendant, opting instead to establish a tiered form of scrutiny based on which some rights could be deemed "fundamental," and therefore subject to searching judicial review,[99] while others were left to a modern rationality review that was judicial review in name only.

This move had earlier roots, but began in earnest with *West Coast Hotel v. Parrish*[100] and *United States v. Carolene*

---

97. *Lochner*, 198 U.S. at 61 (1905) (emphasis added).

98. *Id*. at 64.

99. *Id*.

100. 300 U.S. 379 (1937). Contrary to most popular conceptions, the approach to substantive due process reflected in *Parrish* did not differ materially from that reflected in *Lochner*:

> The Constitution does not speak of freedom of contract. It speaks of liberty and prohibits the deprivation of liberty without due process of law. In prohibiting that deprivation the Constitution does not recognize an absolute and uncontrollable liberty. Liberty in each of its phases has its history and connotation. But the liberty safeguarded is liberty in a social organization which requires the protection of law against the evils which menace the health, safety, morals and welfare of the people. Liberty under the Constitution is thus necessarily subject to the restraints of due process, and regulation which is reasonable in relation to its subject and is adopted in the interests of the community is due process.

Id. at 391. This formulation is broadly consistent with the means-ends scrutiny laid out by the *Lochner* Court, which also required the legislation to have been adopted for a police

Products, Inc, particularly its famous footnote 4,[101] and found its full expression in *Griswold v. Connecticut*.[102] In *Griswold*, the Court considered the continuing reach of *Lochner* and finally issued a clear abrogation of the decision, beginning, "[o]vertones of some arguments suggest that *Lochner v. New York*, 198 U. S. 45, should be our guide. But we decline that invitation . . ."[103] The Court then drew the now-familiar line between fundamental rights that qualify for searching judicial review and other, more quotidian matters, stating, "[w]e do not sit as a super-legislature to determine the wisdom, need, and propriety of laws that touch economic problems, business affairs, or social conditions," before distinguishing the contraception restriction at issue from such laws.[104] This

---

power-related interest and to be reasonably related to that interest. Lochner v. New York, 198 U.S. 45, 56, 61 (1905).

101. 304 U.S. 144 (1938). The text of Footnote 4 is a significant break from *Lochner*'s approach (and I would argue with *Parrish*'s, as well), as it begins the move toward the more categorical and tiered approach to constitutional scrutiny that is dominant today:

> There may be narrower scope for operation of the presumption of constitutionality when legislation appears on its face to be within a specific prohibition of the Constitution, such as those of the first ten amendments, which are deemed equally specific when held to be embraced within the Fourteenth.
>
> It is unnecessary to consider now whether legislation which restricts those political processes which can ordinarily be expected to bring about repeal of undesirable legislation, is to be subjected to more exacting judicial scrutiny under the general prohibitions of the Fourteenth Amendment than are most other types of legislation. On restrictions upon the right to vote; on restraints upon the dissemination of information; on interferences with political organizations; as to prohibition of peaceable assembly.
>
> Nor need we enquire whether similar considerations enter into the review of statutes directed at particular religious or national, or racial minorities: whether prejudice against discrete and insular minorities may be a special condition, which tends seriously to curtail the operation of those political processes ordinarily to be relied upon to protect minorities, and which may call for a correspondingly more searching judicial inquiry.

Id. (citations omitted). But as a decision, it holds that Congress indeed had a rational legislative purpose for regulating filled milk, one reflected in the exhaustive work of legislative staffers and debated in multiple committee hearings, and one that was reflected in the statute itself. *See id*. at 148-49 (reviewing this evidence).

102. 381 U.S. 479 (1965).

103. *Id*. at 481-82.

104. *Id.* at 482. Of course, this was a slight mischaracterization of judicial review during the *Lochner* era, which did not purport to judge the "wisdom" or "propriety" of laws

distinction had the permanent effect of dividing the task of judicial review of laws for constitutionality into two tiers, a task begun in the famous footnote from *Carolene Products*, but cemented into our jurisprudence in *Griswold*.

Under the current tiered form of scrutiny that flows from *Carolene Products* and *Griswold*, legislation that places burdens on fundamental rights will be struck down unless the government can establish that the legislation is narrowly tailored to serve a compelling government interest.[105] In contrast, mere social and economic legislation that does not interfere with a fundamental right will be upheld unless shown to lack "a reasonable relation to a legitimate state interest."[106] This latter test sounds quite similar to the *Lochner* era analysis, certainly placing the burden of proof on the party seeking invalidation, but also seeming to require an actual purpose motivating the legislation, and a rational relationship between the means chosen by the legislature and the purpose it pursues. But in modern application, the standard is dramatically different.

Over time, the New Deal and post-New Deal Courts' disapprovals of the outcome of *Lochner*, but retention of the prospect of judicial review upon a showing of a law's lack of reasonable foundation or relationship to a legitimate end, has morphed into what amounts to judicial abdication or abstention from review entirely in most cases. The current approach had its roots in *Williamson v. Lee Optical*,[107] but found its full expression in *FCC v. Beach Communications, Inc*,[108] an equal protection case that has been understood to articulate the rational basis standard that applies in substantive due process cases, as well.[109]

---

but did ask legislatures to justify the "need" for them to serve a legitimate interest stemming from the police power. *See* Lochner v. New York, 198 U.S. 45, 64 (1905).

105. *See, e.g.*, Plyler v. Doe, 457 U.S. 202, 215-16 (1982) (explaining the levels of scrutiny under the Equal Protection Clause).

106. Washington v. Glucksberg, 521 U.S. 702, 722 (1997).

107. 348 U.S. 483, 487-88 (1955) ("But the law need not be in every respect logically consistent with its aims to be constitutional. It is enough that there is an evil at hand for correction, and that it might be thought that the particular legislative measure was a rational way to correct it.").

108. 508 U.S. 307 (1993).

109. *See, e.g.*, Kelo v. City of New London, 545 U.S. 469, 490 (2005) (Kennedy, J., concurring) (equating the two standards); *Glucksberg*, 521 U.S. at 766 (setting forth the rational basis test for substantive due process).

Justice Thomas, writing for the Court in *FCC*, made clear that the modern approach to rational basis review is not really a doctrine of review at all, but closer to a qualified abstention doctrine.[110] Under this approach, which has been followed in most federal court decisions since, a statute must be upheld under rational basis review "if there is any reasonably conceivable state of facts that could provide a rational basis for the classification."[111] The legislature does not have to proffer such a justification—rather, the Court has the duty to imagine a "conceivable state of facts" that could have motivated the legislature.[112] In addition, "those attacking the rationality of the legislative classification have the burden 'to negative every conceivable basis which might support it.'"[113] Moreover, "it is entirely irrelevant for constitutional purposes whether the conceived reason for the challenged distinction actually motivated the legislature," because the Court "never insisted that a legislative body articulate its reasons for enacting a statute."[114] In fact, "a legislative choice is not subject to courtroom factfinding and may be based on rational speculation unsupported by evidence or empirical data."[115]

Thus, under the current approach the courts have developed, the court must uphold the legislation unless the plaintiff invalidates any and all conceivable justifications for the means chosen, whether real or imagined. The government does not bear any burden of production or persuasion as to either the ends it seeks to serve or the means it has chosen to serve such ends.[116] This extreme deference that courts now give to most legislative enactments requires neither fact nor logic to sustain their rationality—it even indulges judicial speculation of what a legislature "could rationally believe,"[117] without any

110. *FCC*, 508 U.S. at 313-14.

111. *Id*. at 313.

112. *Id*.

113. *Id*. at 315 (quoting Lehnhausen v. Lake Shore Auto Parts Co., 410 U.S. 356, 364 (1973)).

114. *Id*. (citing United States R.R. Retirement Bd. v. Fritz, 449 U.S. 166, 179 (1980)).

115. *Id*.

116. Clark Neily, *No Such Thing: Litigating under the Rational Basis Test*, 1 N.Y.U. J.L. & LIBERTY 897, 912 (2005).

117. Cook v. Stewart, 28 F. Supp. 3d 1207, 1213 (N.D. Fla. 2014).

requirement to even found that speculation in evidence of which the legislature would have been aware at the time.

This regime, in its application if not its rhetoric, is obviously a far cry from that of the *Lochner* era, and even, I would argue, the era in which *Lochner* was initially disapproved, and then rejected.[118] It is also inconsistent with any concept of judicial review stretching beyond abstention, and it is worth asking whether we made a wrong turn in moving quite so far away from that era's constitutional norms. The following section proposes a potential correction, focusing on the salutary features of *Lochner*-era jurisprudence that need not have been left behind in the effort to reject its problematic implications for labor law and the New Deal.

IV. Reviving Rational Basis Review

The extreme, yet real and recent, example from Florida above illustrates that rational basis review of substantive due process claims has become little more than judicial abdication—something more akin to the political question doctrine than an actual doctrine of merits review. This Part makes the case for a way forward, which turns out to be a way backward in Constitutional law.

The examples of the absurdity of results that the current approach to rational basis review produces are legion. Many have been outlined in the careful work of constitutional lawyer Clark Neily,[119] and others have been catalogued over the years in other scholarship.[120] Although not highlighted much in discussions of the decision, Justice Marshall's liberal dissent to *San Antonio v. Rodriguez* also spends significant time criticizing the too-lenient approach of the majority in reviewing Texas's property-tax-based school funding program for rationality.[121] In particular, Justice Marshall outlines the lack of factual

118. *See supra* notes 100-04 and accompanying text (discussing *Parrish*, *Carolene Products*, and *Griswold*).

119. *E.g.*, Neily, *supra* note 116, at 903-13.

120. *E.g.*, Jeffrey D. Jackson, *Classical Rational Basis and the Right to be Free of Arbitrary Legislation*, 14 GEO. J. L. & PUB. POL'Y 493, 503 (2016).

121. San Antonio Ind. Sch. Dist. v. Rodriguez, 411 U.S. 1, 98-99 (1973) (Marshall, J., dissenting) (criticizing rigid tiered scrutiny, and arguing for more of a sliding scale of review).

foundation or logic underlying Texas's purported justification of preserving "local control" in relation to its chosen means of funding education primarily through local property taxation, which creates substantial inequalities in funding throughout the state's school districts, leaving wealthy districts with enough funding to truly exercise local control, while saddling poor districts with barely enough money to make a decent effort at meeting the most basic of state standards.[122] In choosing to fight on this ground and highlight the internal contradictions in Texas's school funding plan, Justice Marshall draws substantially from the pre-*FCC*, and even pre-*Griswold*, approach to rational basis review.

The decades-long move away from the *Lochner* era was a well-intentioned one, which sought to preserve both the expansions in worker protections that both preceded and followed the Great Depression and to forestall opportunistic challenges to New Deal and Great Society legislation. But it overshot its mark. The proof of this overshooting lies in decisions that nominally fell under the nearly absolute deference-based standard articulated in *FCC*, but which nevertheless came in for far more searching judicial review.

Beginning with *Plyler v. Doe*,[123] the Court confronted a case in which it had made clear less than a decade prior that strict judicial scrutiny did not apply.[124] In *Plyler*, the statute under challenge denied any public educational services to undocumented immigrants residing in Texas.[125] The Court quickly reaffirmed its holding in *Rodriguez* that education was not a fundamental right for the purpose of due process or equal protection analysis, but also made the point that it was not equivalent to any ordinary social benefit either.[126] It also held that undocumented immigration status could not be treated as a suspect classification.[127] These two holdings should have shunted the case into rational basis review territory. And rhetorically, it did, as the Court referred to "rationality" in

122. *Id*. at 126-28.
123. 457 U.S. 202 (1982).
124. *Rodriguez*, 411 U.S. 1, 26.
125. *Plyler*, 457 U.S. at 205.
126. *Id*. at 221.
127. *Id*. at 223.

preparing to conduct its review, but the Court also laid out a particularly searching form of rationality review—one more reminiscent of *Lochner* than *FCC*: "[i]n determining the rationality of § 21.031, we may appropriately take into account its costs to the Nation and to the innocent children who are its victims."[128]

Upon a skeptical review, the Court rejected purported justifications for the law based on national immigration policy;[129] conservation of scarce state resources;[130] preventing an influx of undocumented immigrants into the state;[131] the special burdens that undocumented immigrant children place on state educational delivery;[132] and a lack of expected benefit to the state due to the tendency of migrants to move around the country.[133] In most cases, the Court rejected these justifications due to their lack of evidentiary support in the record.[134] In short, the Court, even though it applied rational basis review, actually engaged in judicial review. Despite prior judicial protestations to the contrary,[135] the sky did not fall, and the legislature adjusted, and then went right on legislating.

The Court went even further in *City of Cleburne v. Cleburne Living Center, Inc.*,[136] rejecting the idea of a quasi-suspect classification for the mentally disabled, and holding that rational basis review would apply to a local ordinance placing significant burdens on obtaining a permit to build a living center for those with mental disabilities.[137] But even though the Court selected this highly-deferential standard, it nevertheless upheld the lower court's decision striking down the ordinance, stating, "[b]ecause in our view the record does not reveal any rational basis for believing that the Featherston home would pose any

128. *Id*. at 224.
129. *Id*. at 226.
130. *Plyer*, 457 U.S. at 227.
131. *Id*. at 228.
132. *Id*. at 229.
133. *Id*. at 230.
134. *Id*. at 224-30.
135. FCC v. Beach Commc'ns, Inc., 508 U.S. at 315 (quoting Lehnhausen v. Lake Shore Auto Parts Co., 410 U.S. 356, 365 (1973) ("Only by faithful adherence to this guiding principle of judicial review of legislation is it possible to preserve to the legislative branch its rightful independence and its ability to function.")).
136. 473 U.S. 432 (1985).
137. *Id*. at 446-47.

special threat to the city's legitimate interests, we affirm the judgment below insofar as it holds the ordinance invalid as applied in this case."[138]

Rather than imagining a rational basis that "could have" motivated the adoption of the ordinance, the Court reviewed what the city claimed actually motivated its adoption, and found itself not convinced.[139] The Court ultimately held:

> "The short of it is that requiring the permit in this case appears to us to rest on an irrational prejudice against the mentally retarded, including those who would occupy the Featherston facility and who would live under the closely supervised and highly regulated conditions expressly provided for by state and federal law."[140]

Rejecting each of five different justifications the city offered for the ordinance, the Court illustrated that judicial skepticism of even claims of legitimate governmental ends is warranted in at least some cases, even where fundamental rights and suspect classifications are not at issue.

Perhaps these decisions can be explained by the fact that they both pre-dated *FCC* and its articulation of the extremely deferential approach in an authoritative way, or perhaps because they were both equal protection cases, rather than substantive due process cases. But *FCC* did not make any effort to abrogate or overrule these cases, and the rational basis standard has long been applied coextensively and consistently between equal protection and substantive due process cases. They were also followed by other rulings that nominally fell into the rational basis category but wound up applying a more *Lochner*-like form of rational basis review than one would have expected with FCC on the books.[141]

---

138. *Id*. at 448.

139. *Id*. at 448-50.

140. *Id*. at 450.

141. *E.g.*, Romer v. Evans, 517 U.S. 620, 632-35 (1996) (applying rational basis review to invalidate a Colorado constitutional amendment for irrationally imposing legislative disabilities on gay men, lesbians, and bisexuals in local government); Quinn v. Milsap, 491 U.S. 95, 109 (1989) (invalidating a property ownership requirement to sit on a county land use board as an irrational classification). For a detailed treatment of all Supreme Court cases since 1971 in which the Court has applied rational basis review in a searching way more reminiscent of *Lochner* than *FCC*, see Note, Raphael Holoszyc-

Based on the discussion above, the level of judicial deference reflected in *FCC*, and in the Florida suit challenging aberrant uses of value-added modeling, is neither obligatory on the courts nor uniquely preservative of the separation of powers. This means that there is no real principled basis to consign the substantive due process method of the *Lochner* era to the trash heap. Rather, we should reconsider at least requiring some level of means-ends connection, presumed to exist absent a challenge, but able to be put at issue by a plaintiff bearing a burden of proof. Only this burden should be one actually possible to meet through documentable fact, expert testimony, legislative history, etc., that negatives the connection between the asserted goal and the means chosen to meet it, or through some other showing that this ends-means relationship was not actually the basis for a challenged piece of legislation. Had this opportunity been available to the plaintiffs in *Cook v. Stewart*, it is likely that at least one, and perhaps both, groups of teachers would have prevailed. The plaintiffs in *San Antonio v. Rodriguez* would have had a road to victory, as well, not to mention the scores of individuals seeking to operate businesses that present no danger to the public, and who are burdened by restrictions on their right to earn a living that have no connection to public health, safety, or welfare.[142]

I have argued in the past that, under both the United States Constitution and every state constitution, government stands in a fiduciary capacity in relation to the people.[143] Others have also claimed that, in our republican form of government, we delegate to our elected officials the power to act on our behalf, and in accepting that delegation and power, they assume the duty to act in our best interests, and to do so both faithfully and rationally.[144]

Pimentel, *Reconciling Rational Basis Review: When Does Rational Basis Bite*, 90 N.Y.U. L. REV. 2070, 2106-17 (2015).

142. For a thorough sampling of such individuals, see Neily, *supra* note 116.

143. Scott R. Bauries, *A Common Law Constitutionalism for the Right to Education*, 48 GA. L. REV. 949, 986-87 (2014); Scott R. Bauries, *The Education Duty*, 47 WAKE FOREST L. REV. 705 (2012) [hereinafter *Education Duty*].

144. *See* Gary Lawson & Guy I. Seidman, *By Any Other Name: Rational Basis Inquiry and the Federal Government's Duty of Care*, 69 FLA. L. REV. 1385, 1405 (2017) (arguing for fiduciary duty as the basis of rational basis review); Sotirios A. Barber, *Are Professors Lawson and Seidman Serious about a "Fiduciary Constitution"?*, 69 FLA. L. REV. F. 10, 11 (2017), http://www.floridalawreview.com/issue/volume-69/

Some who agree with this claim place rational basis review within the duty of due care, on the theory that fiduciaries assume a baseline duty to act rationally as to their entrusted work.[145] This conception would seem to support something more than the rational basis review of today—but not much more. The duty of care in other fiduciary contexts is typically enforced through some form of process-based review, with the overall rationality of the decision not being directly questioned. Accordingly, I see the duty of care as more fitted to the limited circumstances in which a legislature or other government actor assumes a positive duty to act, such as under state constitutional education clauses.[146] My own work places all individual negative rights enforcement under the duty of loyalty.[147] In the absence of specified duties to legislate on particular matters, such as education in the states, legislative duty stems from its discretionary power to legislate—or not legislate. But understanding that legislative positions are delegations of authority from the public, not patronages or sinecures for personal enrichment, means that legislatures cannot legislate beyond the background basis for their delegated power. Underlying this power is the background duty to legislate in the best interests of the entrusting public—to have a public purpose for legislating, and to legislate in a way that is directed is serving that purpose.

This bedrock requirement is what forms the basis of the "police power"—the starting point of all rational basis review of state action during the *Lochner* era. An act that either legislates outside of the police power, or uses the police power oppressively, pretextually, or in a self-dealing way, is therefore an act that is disloyal to the beneficiary of the legislative duty—the public. If this is so, and it seems abundantly clear in both the

---

[https://perma.cc/KDR6-QRHX] ("Indeed, a fiduciary constitution would seem to compel substantive reasonableness for any governmental act. No mentally competent person would voluntarily delegate power to an agent to be exercised carelessly or pretextually or for anything less than an understanding and reasonably competent pursuit of the principal's interest.").

145. Lawson & Seidman, *supra* note 144, at 1404-07.

146. *Education Duty*, *supra* note 143, at 747-48.

147. *Education Duty*, *supra* note 143, at 747-48.

founding documents and their influences,[148] then the courts are on sound footing reviewing legislation for substantive rationality. Indeed, given their role and their co-equal status as public fiduciaries, the courts themselves likely have their own fiduciary duties to engage in such review, and to do so in a searching way.[149] Reviewing—actually reviewing—legislation for whether it is rationally directed to serve a proper legislative purpose is therefore the proper and legitimate role for the courts, one they have abdicated over time by gradually ratcheting down the standards for legislative rationality.

Conclusion

Some may view the current practice of rational basis review as a correct reflection of the extreme deference that courts should afford legislative acts, exemplified by the common state constitutional law refrain that legislation should not be invalidated unless its unconstitutionality is shown "beyond a reasonable doubt."[150] Others might posit that the staying power of rational basis review, as practiced in the modern era, is justified by a fear of re-*Lochner*ing the Constitution.[151] But we do not have to love *Lochner* itself to seek to restore its mode of rational basis review.

My project here has not been to apologize for or defend the outcome of the *Lochner* decision itself. I have my own views as

148. *See, e.g.*, EVAN FOX-DECENT, SOVEREIGNTY'S PROMISE: THE STATE AS FIDUCIARY 28-51 (2011); GARY LAWSON ET AL., THE ORIGINS OF THE NECESSARY AND PROPER CLAUSE 56–57 (2010); Evan Fox-Decent, *The Fiduciary Nature of State Legal Authority*, 31 QUEEN'S L.J. 259, 260–61 (2005); Evan J. Criddle, *Fiduciary Foundations of Administrative Law*, 54 UCLA L. REV. 117, 120 (2006); Sung Hui Kim, *The Last Temptation of Congress: Legislator Insider Trading and the Fiduciary Norm Against Corruption*, 98 CORNELL L. REV. 845, 903-04 (2013); D. Theodore Rave, *Politicians as Fiduciaries*, 126 HARV. L. REV. 671, 677 (2013).

149. *See* Ethan J. Leib, David L. Ponet & Michael Serota, *A Fiduciary Theory of Judging*, 101 CALIF. L. REV. 699, 714 (2013) (identifying the fiduciary conception of judging underlying the founding documents, establishing the duty to, among other things, "keep[] the legislature within its bounded authority").

150. *See* Scott R. Bauries, *State Constitutions and Individual Rights: Conceptual Convergence in School Finance Litigation*, 18 GEO. MASON L. REV. 302, 358 (2011) (describing the "beyond a reasonable doubt" standard of constitutionality).

151. *See* David A. Strauss, *Why Was Lochner Wrong?*, 70 U. CHI. L. REV. 373, 373-86 (2003) (carefully working through the arguments as to why *Lochner* was wrongly decided, in an effort to harmonize its rejection with current constitutional law).

to why it was wrongly decided, and those can be better expressed in an article more focused on that topic. My concern here has been with the methodology of *Lochner*, not its result. Using an exemplar case coming out of the education context, I have sought to show how, in rejecting a decision that we can probably all agree came out the wrong way, we also rejected the public-protective, and therefore proper, role for the courts as a check on legislative action that is irrational, protectionist, rent-enabling, harmful to public servants, or otherwise contrary to the public interest—one that has worked well in many cases other than *Lochner*.[152] Correcting that error is as easy as privileging precedents such as *Plyler* and *Cleburne* over those such as *FCC*. It's high time we considered that.

152. *See supra* notes 100-04 and accompanying text (discussing *Plyler*, *Cleburne*, and *Lawrence*). For an illuminating historical discussion of the *Lochner* era cases that upheld public protective legislation—even workplace restrictions—against constitutional challenges, see DAVID E. BERNSTEIN, REHABILITATING *LOCHNER*: DEFENDING INDIVIDUAL RIGHTS AGAINST PROGRESSIVE REFORM 51 (U. Chi. Press 2011).